SUMMARY

FUNCTIONAL GENOMICS AND ENVIRONMENTAL HEALTH

Sponsored by

Division of Bioengineering and Environmental Health
Massachusetts Institute of Technology

and the

National Institute of Environmental Health Sciences
National Center for Toxicogenomics

December 11, 2000

Summary of Workshop on
Functional Genomics and Environmental Health

December 11, 2000
Massachusetts Institute of Technology

Introduction

The National Institute of Environmental Health Sciences (NIEHS) established the National Center for Toxicogenomics (NCT) in June 2000. Toxicogenomics is a scientific field that combines studies of genetics, genomic-scale, mRNA expression, cell and tissue-wide protein expression, and bioinformatics to understand the roles of gene-environment interactions in disease. The NCT was created to facilitate application of toxicogenomics to improve human health.

The goals of NCT are: 1) To help facilitate the use of gene expression profiling, proteomics and bioinformatics to the understanding of how cells respond to stress; 2) to create a public database relating environmental stress to biological responses; 3) to collect information relating environmental exposure to disease; 4) to develop an improved paradigm for use of computational mathematics for understanding response to environmental stress; and 5) to identify biomarkers of disease or exposure to enhance environmental health.

To help define the path toward fulfilling its goals, a series of workshops are being held through which the research community can learn about the NCT and provide input to the NCT. The first workshop in this series entitled "Functional Genomics and Environmental Health" was held at the Massachusetts Institute of Technology (MIT) in Cambridge, MA and was co-organized by Ben Van Houten (NIEHS) and Leona Samson (Harvard School of Public Health).

In his opening remarks, Samuel Wilson, Deputy Director (NIEHS) spelled out the goals of this meeting. The scientific community is excited about the potential of techniques such as DNA microarray expression profiling to advance knowledge of basic biology and to improve human health and medical practice. The fast pace and continually changing new field of toxicogenomics presents many opportunities and challenges to the research community; however, the future of the field is not well defined at present. We are now faced with the challenge of lending definition to the field and understanding where the field of toxicogenomics is going. One endpoint of this workshop, and the NCT program itself, is to evaluate "problem definition" in understanding the cellular response to stress. In addition, this workshop will begin to define how toxicogenomics might change the practice of toxicology research.

The workshop included scientific presentations by Richard A. Young (Whitehead Institute), George Church (Harvard Medical School), Edwin Clark (Millennium Pharmaceuticals), and Leona Samson, all of whom have extensive experience in applying DNA microarray technology to problems of basic biology. The scientific presentations were followed by a roundtable discussion session. This document summarizes the main points of each presentation and of the discussion session.

Transcriptional Networks in Yeast and Human Macrophages

Richard Young presented information on three topics: hardware and software issues for DNA microarray studies, microarray studies of transcription networks in yeast, and microarray studies of primary human macrophages.

Young described several types of microarrays currently in use that are distinguished from one another by their method of preparation and their source. Commercial microarrays are produced by companies including Affymetrix, Incyte, Agilent and Corning. These arrays are prepared either by in situ oligonucleotide synthesis of probes or by spotting probes onto a glass slide surface. When compared directly, the commercially available arrays appear to be of similar quality. In contrast, Young indicated that "homemade" glass slide arrays, prepared in many facilities where microarray research is being carried out, are subject to large variations in quality and can be of much lower quality than an equivalent commercially-prepared array.

The methods for analyzing microarray data are evolving rapidly. Scatter plots are commonly used at present to illustrate and present microarray data. The scatter plot compares data obtained when a single sample is differentially labeled with two dyes. For a majority of investigators it has been customary to determine significance of a gene expression change using an arbitrary level such as 2-fold or higher. Young indicated that this should no longer the method of choice for estimating significance in microarray data. A method was recently developed in which an error model is used to determine an error probability boundary. Error models generate a p-value for each data point. This method is valuable because it allows significance to be determined with statistical estimates of confidence and error. Young emphasized that this method is far superior to more commonly used methods, and he encouraged researchers to begin to use it more widely.

Young indicated that database management is a key component of microarray studies. One database system called "Resolver" has been developed by Rosetta Inpharmatics. This system is available from Rosetta, but for many microarray researchers it may be prohibitively expensive.

Transcriptional networks, especially in yeast and eukaryotic organisms, are a major research interest of Young's. The mechanism of transcription is complex. Other complex processes are also involved in transcriptional networks such as the following: the interaction between transcription factors and chromatin; modulation of chromatin structure; coordination and recruitment of transcription cofactors; and regulatory mechanisms that co-regulate groups of genes. Transcriptional networks for metabolic pathways have been elucidated in great detail over the past several decades. The understanding of these pathways is used extensively in drug development. Thus, one rationale for studying transcriptional networks is to enhance and expedite the drug discovery process.

Young and his colleagues have concluded that simple microarray approaches can not be used to deduce transcriptional networks. For example, they studied the expression of 6218 yeast genes in cells exposed to hydrogen peroxide. Expression profiles showed that approximately 30% of the yeast genome was involved in a short-term transient response to hydrogen peroxide. Attempts were made to discern which of these responses were linked in a significant manner using iterative simulations and neural network theory, but these efforts were not successful.

To better understand transcriptional networks, Young developed an approach called genome-wide location analysis (GWLA). The first step in GWLA is to crosslink a crude cell extract using a reagent that forms DNA-protein and protein/protein crosslinks. The DNA is sheared into short fragments, differentially labeled, and then screened with a selective probe such as a monoclonal antibody. Two differentially labeled DNA pools result; one pool is enriched for a bound protein and one is not enriched. These DNA pools are hybridized to an array containing all intergenic regions of the yeast genome. The magnitude of the signal is compared in enriched and unenriched DNA; a potential network of protein-DNA interactions can then be deduced.

Genome-wide location analysis was applied to the transcriptional networks of the very well studied yeast transcription factor *Gal4*. The goal of this GWLA experiment was to identify genes that are bound and upregulated by *Gal4* in the presence of galactose. Ten genes were identified; 7 of the genes are known targets of *Gal4* and 3 genes were not previously known to interact with *Gal4*. The new *Gal4* targets are *FUR4, PCL10* and *MTH1*, which are genes involved in sugar metabolism in yeast. In addition, the algorithm AlignACE identified a putative *Gal4* binding motif in the upstream region of these genes.

Young emphasized his view that GWLA is a powerful tool for understanding transcriptional networks. He hopes to apply this technique systematically to up to 200 transcriptional activators and repressors in yeast, and he eventually hopes to use GWLA to analyze gene regulation in higher eukaryotic cells. In the future, GWLA may help annotate the function of many genes and improve our understanding of complex pathways such as genome replication and repair.

Young has also used microarray analysis to understand the response of human macrophages to infectious agents. Macrophages in culture were exposed to bacteria and harvested after different periods of exposure. Extracts of exposed macrophages were analyzed using a high density Affymetrix array for expression of 2500 human genes. Cluster analysis revealed that exposure to 8 infectious agents, including strains of *Escherichia coli*, Salmonella, Staphylococcus, *Mycobacterium tuberculosis* (TB) and others, induced a similar group of genes. Analysis with an array for 68 human cytokine and chemokine genes revealed that about a fourth of these genes were induced in a similar manner after exposure to different bacteria. Further analysis indicated that bacterial lipopolysaccharide and heat shock proteins were likely to be inducing agents contributing to this pattern of macrophage gene expression. Careful study identified a unique characteristic of the response to TB. Macrophages fail to induce interleukin-12 after exposure to TB or after combined exposure to TB and *E. coli*. However, cells exposed only to *E. coli* induce interleukin-12 expression. The results indicate that IL-12 therapy might be useful to enhance recovery from TB infection; importantly, this therapy might be useful in cases involving a drug-resistant strain of TB.

Methods to Advance Toxicogenomics

George Church is interested in understanding cells by quantifying all cellular functions and components and their interrelationships. Church described several techniques developed by his laboratory that may be useful to researchers in the field of toxicogenomics.

Church has developed an in situ method to amplify single DNA molecules. DNA is fixed to a glass surface, embedded in a gel matrix and amplified in situ by PCR. This approach uses PCR

to create a slide with DNA spots also called PCR colonies or polonies. Polonies can be sequenced by iterative DNA synthesis cycles using fluorescent dyes. Polonies might be used to study gene polymorphism, prepare recombinant DNA, isolate specific DNA fragments, carry out other recombinant DNA procedures or to prepare DNA arrays.

Church has also developed a high resolution genome array for study of gene expression in *E. coli*. This array has on average one 25 bp probe for every 30 bp covering the entire *E. coli* genome. Each gene is represented by multiple probes, overcoming the effects of poor hybridizing probes or secondary structure which decrease hybridization efficiency. Because the array has comprehensive representation of both intergenic and genic sequences, it has unusually high sensitivity and can detect very low abundance transcripts.

Computational methods to analyze DNA sequences are important to interpret microarray data. Another method developed by Church is an algorithm called AlignACE which finds common patterns in DNA sequences. For example, AlignACE can be used to find a putative transcription factor binding site in the promoters of a set of coregulated genes identified by expression profiling.

Metastasis and Pharmacogenomics

Edwin Clark described studies of metastasis and breast cancer drug efficacy using microarray technology. Metastatic cancer is the principal cause of cancer deaths. Clark analyzed the genetic basis for metastatic potential starting with the poorly metastatic human A375P melanoma cell line. A375P cells were injected into nude mice and allowed to form tumor nodules in the lungs of the injected mice. Cells from the metastasized tumor nodules were isolated, established in culture and then the cycle was repeated. This selection process produced a series of increasingly metastatic derivatives of the A375P parental cell line, and a highly metastatic variant called A375M.

Gene expression was compared in the metastatic variant cells and the parental A375P cells using an Affymetrix-type microarray with 6000 human genes. Many genes involved in the cytoskeleton were expressed at a higher level in A375M and other variant cells than in A375P cells. Fibronectin, RhoC and thymosinβ4 were the most highly induced genes (>10-fold induction) in the A375M cell line. Clark investigated the relationship between RhoC and metastatic potential in more detail. For example, retroviral particles were prepared containing recombinant wild type RhoC or a dominant negative mutant of RhoC. A375P cells were infected with the particles and injected into nude mice. RhoC overexpression significantly increased metastatic potential and the dominant negative mutant of RhoC significantly decreased metastatic potential. Clark concluded that the function of RhoC, which increases cellular motility, is essential for cells to metastasize. It is possible that RhoC could be developed as a diagnostic tool of metastatic potential or as a drug target to inhibit metastasis.

Microarray research has great potential to accelerate progress in the field of pharmacogenomics. The goal of pharmacogenomics is to understand the factors that determine drug efficacy and safety for each individual patient. Ultimately, this kind of understanding will allow medical professionals to predict drug responsiveness and customize medical treatment for each patient. Clearly, this knowledge will help produce the best medical outcome in each patient.

Clark studied the efficacy of taxol or taxol combination therapy to treat ovarian cancer. A group of 51 patients who received chemotherapy were tracked for 5 years. Twenty-seven patients relapsed (nonresponders) and 24 patients remained disease free (responders). Gene expression analyses were carried out with a 36000 gene microarray comparing responders and nonresponders. A very large number of genes were selectively induced or repressed in the responders relative to the nonresponders. A group of proprietary and nonproprietary marker selection algorithms (*i.e.*, ET, Ebayes, Ecombo, POOF, Class Predictor) were used to analyze the data. The analysis identified 8 genes which have differential expression patterns strongly associated with drug efficacy. Using this set of genes as diagnostic markers, the expression pattern of an individual patient correctly predicted outcome of drug treatment in 43 of the 51 cases. Clark is currently testing if similar success is achieved when the 8 gene diagnostic expression profile is used to predict treatment outcome for a larger group of ovarian cancer patients already receiving treatment. An in situ histochemistry test for expression of 5 genes has also been developed based on this analysis. These results indicate that microarray studies are having a large impact on the field of pharmacogenomics.

Transcriptional Regulation of the Response to DNA Damage in Yeast

Leona Samson described a relatively focused set of experiments using Affymetrix-based microarrays to study DNA repair in yeast. Samson's research is focused on the cellular response to DNA damage caused by DNA alkylating agents. Exposure to DNA alkylating agents leads to altered forms of the DNA bases called adducts. The presence of alkylation adducts in DNA can lead to gene dysfunction and disease, especially when the DNA adducts are not repaired. Several DNA repair pathways have been characterized, including nucleotide excision repair, base excision repair, mismatch repair and recombinational repair. Base excision repair is the primary pathway that recognizes and repairs alkylation damage, although there is considerable overlap in the specificity of most DNA repair pathways. Glycosylases are important enzymes in the repair of DNA alkylation adducts, because they carry out the recognition step during adduct repair and excise the adducts from DNA. In earlier studies, Samson's group characterized this group of enzymes and cloned their genes. Wild-type and mutant forms of the proteins have been overexpressed and studied.

The eukaryotic 3-methyl adenine glycosylase (MAG) is a homologue of the well characterized *E. coli* enzyme AlkA, which is an important player in the response to reagents such as methyl methanesulfonate (MMS) and N-methyl-N'-nitro-N-nitrosoguanidine (MNNG).

An expression profiling study was carried out in yeast to determine the global response to alkylating agents. Affymetrix chips were used that monitor expression in all known yeast genes (approximately 6000 genes) using approximately 20 oligonucleotide probes per gene. Yeast cells were treated with and without MMS (methyl methane-sulfonate), RNA isolated, cDNA prepared and the cDNA was hybridized to chips. A large number of inducible genes were found: 294 genes were upregulated 4 to 230-fold, and 139 genes were down-regulated significantly. The expression pattern of about 5000 yeast genes did not change in response to MMS.

The data obtained by microarray was validated by Northern blot experiments. Correspondence for the two methods was very good; the two data sets have a correlation coefficient of 0.8.

The MMS-inducible genes in yeast had many different cellular functions, which included the following: stress response, DNA repair, DNA replication, cell cycle, cell signaling, cell wall, cell transport, sulfur metabolism, mRNA metabolism, RNA transcription, protein secretion, cytoskeleton, chromatin structure, amino acid metabolism, protein degradation and others, including genes of unknown function. The latter category included a large proportion of the MMS-induced genes. Genes involved in amino acid metabolism and protein degradation were apparently disproportionately induced in yeast cells treated with MMS. The MMS-repressible genes also included many genes involved in protein metabolism, especially genes for ribosomal proteins.

Additional parameters were examined that influence the expression profile in these cells. A total of 26 different experimental conditions were tested, such as MMS dose, kinetics of transcriptional response, cell cycle at time of MMS treatment, and other alkylating or DNA damaging agents including MNNG (methyl nitro-nitroso-guarudine), t-butyl hydrogen peroxide, γ-irradiation, 1,3-Bis[2-chloroethyl]-1-nitrosourea and 4-nitroquinoline-1-oxide. Twenty-one genes were found that respond similarly to all agents tested. These data were subject to cluster analysis and 18"self-organized maps" or clusters were defined. One of these clusters includes yeast MAG glycosylase and 212 other genes.

The cluster containing MAG includes many genes involved in protein degradation. All of these genes carry an upstream regulatory sequence (URS2) with the sequence GGTGGCGA, which is similar to the proteosome associated control element GGTGGCAA. Furthermore, the yeast protein *RPN4* binds this sequence and regulates expression from these genes. *RPN4* is a protein degradation gene and a transcriptional activator. *RPN4* deficient (rpn4) yeast are sensitive to DNA damage induced by MMS and other DNA damaging agents. In addition, expression analysis of a rpn4 strain indicates a very different pattern than wild type yeast in the absence and presence of MMS. Thus, these experiments describe the use of expression profiling to identify a novel and potentially important regulator of the stress response in yeast.

Roundtable Discussion: Opportunities and Challenges in Toxicogenomics

Ray Tennant (Director, NCT, NIEHS) briefly addressed some of the goals and challenges facing NCT. Research in toxicogenomics has the potential to help solve basic problems of environmental health, including identifying, classifying and assessing hazards, assessing exposure, and identifying susceptibility factors. However, it remains unclear how the methods of toxicogenomics can best be applied to these issues, and a great deal of future research will be required. Partnerships and collaborations are needed on a broad level. Thus, NCT seeks to establish a broad consortium of academic and private sector research groups working in toxicogenomics. The initial goal of NCT is to establish a toxicogenomics database (Chemical Effects on Biological Systems Database, CEBS-DB), which will serve the larger toxicogenomics research community.

Ben Van Houten (Extramural Coordinator, NCT) mentioned that NIEHS released a request for application (RFA) to form a Toxicogenomics Research Consortium (TCR) in November, 2000. Michael McClure (NIEHS) is the contact for information regarding this RFA. Van Houten briefly described the structure of the TRC, which is a major extramural initiative within NCT. The TCR will consist of academic member groups (initially 5-6), a central information

technology contractor who will provide gene expression profiling support and establish and maintain the CEBS database, and supporting NIEHS staff and scientific advisors from the extramural and intramural NIEHS programs.

The present RFA is to develop a national TRC that will increase the capacity of the extramural research community to apply microarray gene expression profiling to the understanding of biological responses to environmental stress. This will be accomplished by: 1) accelerating research in the area of environmental stress response using microarray gene expression profiling, 2) developing standards and practices that will allow analysis of gene expression data across platforms and provide intra- and inter laboratory validation, and 3) contribute to the development of a relational database for gene expression data.

Van Houten also mentioned several important upcoming NCT meetings: January 24, 2001, Proteomics Workshop, Tucson, AZ; February 2, 2001, pre-application meeting, NIEHS, Research Triangle Park, NC; March 5, 2001, Bioinformatics Workshop, North Carolina State University, Raleigh, NC.

Ben Van Houten and Richard Paules (Co-Director, NIEHS Microarray Center, NIEHS) opened the discussion session of this workshop by presenting the discussants with the following set of questions:

What areas in functional genomics can best or only be advanced through the collective efforts of a research consortium and what areas in functional genomics can best or only be advanced outside of a consortium?

- How can we best facilitate the coordinated efforts of leading scientists toward accomplishing an agreed upon goal while encouraging and supporting creative and independent research?
- What are the critical issues that will need to be addressed to ensure the success of a consortium effort?
- How can we involve scientists who are not officially consortium members?
- What are the unique opportunities in toxicology created by recent advances in genomics that will define the area of toxicogenomics?

There was much valuable discussion on these issues. The following bullets and paragraphs attempt to capture some of the key discussion points emerging from this session.

- The discussants explored the pros and cons of establishing a narrow focus for the NCT consortium. Most of the discussants expressed the feeling that the NCT needs to establish clearly defined goals in order to move forward. There was significant debate over how to do this, which goals might be achievable and which might not, and which goals are strategically optimal. Specific model systems were suggested for microarray research including yeast, mouse liver, or *Caenorhabditis elegans*. It was also suggested that the NCT should select a limited number of toxicants for study. It was pointed out that yeast provide an advantage due to availability of knockouts for all yeast genes.

- Several discussants suggested that NCT should develop a strategy akin to a business plan. To do this, NCT needs to establish clear goals, estimate the resources needed to those goals, and compare that estimate with the resources that are available. Some concern was expressed that available funds are limiting and may not be sufficient to achieve some of the goals of NCT. For example, it may not be feasible for NCT researchers to use commercial arrays, even if they are presently the highest quality arrays available. To function well, NCT will also need to develop ways to evaluate its progress and success.

- Samson, Graham Walker (MIT), Katherine Dixon (University of Cincinnati) and many others encouraged NCT to maintain an emphasis on mechanistic research and on the biological relevance of projects undertaken by NCT.

- There was agreement that NCT should target researchers in the toxicology community as well as other outstanding researchers who are not currently working in toxicology.

- The field of microarray research is still in a great deal of flux. This in some ways limits what NCT might be able to envision in the near future. For example, Young mentioned that he in some cases feels it necessary to discard the results of a study because of advances made over the last 6 months to a year. He also felt the field is moving so rapidly that the present status of the field would be completely outdated within 5 years. Young encouraged the NCT leadership to maintain a flexible agenda and to evaluate their progress and goals at least every 6 months.

- Young, Clark, Rowan Chapman (Rosetta Inpharmatics, Inc.), Steve Tannenbaum (MIT), Larry Marnet (Vanderbilt) and others clearly voiced the feeling that NCT could help the microarray field by establishing standards and protocols for this research community. This effort is essential in order for NCT to function, but it also would help increase stability in the field by promoting uniformity within this research community. Chapman described Rosetta's efforts that contribute to this goal. Rosetta has been involved in the development of a uniform data format exchange called Gene Expression Markup Language (GEML™). GEML is available to the public at the URL htpp://www.GEML.org.

- However, considerable concern was expressed about the validity of comparing data generated with different platforms. This is one of the main goals of the TRC, and it is clear that cross-platform validation studies must be performed within the consortium. Some discussants were optimistic that this goal can be achieved via the NCT plan as outlined by Van Houten.

- Marnet expressed concern that small research groups, or groups not yet engaged in microarray analysis, would not be able to participate in the NCT. He urged the NCT to try to be more inclusive than exclusive in this regard, and not to limit participation to the research groups identified in the initial NCT plan. The discussions concluded that this could be achieved by providing standard platform, once developed, to all research groups. There was also important discussion on considerations for provision of training aspects for small research groups – this should be included.